

Multilevel regression mixture analysis

Bengt Muthén

University of California, Los Angeles, USA

and Tihomir Asparouhov

Muthén & Muthén, Los Angeles, USA

[Received January 2008. Final revision September 2008]

Summary. A two-level regression mixture model is discussed and contrasted with the conventional two-level regression model. Simulated and real data shed light on the modelling alternatives. The real data analyses investigate gender differences in mathematics achievement from the US National Education Longitudinal Survey. The two-level regression mixture analyses show that unobserved heterogeneity should not be presupposed to exist only at level 2 at the expense of level 1. Both the simulated and the real data analyses show that level 1 heterogeneity in the form of latent classes can be mistaken for level 2 heterogeneity in the form of the random effects that are used in conventional two-level regression analysis. Because of this, mixture models have an important role to play in multilevel regression analyses. Mixture models allow heterogeneity to be investigated more fully, more correctly attributing different portions of the heterogeneity to the different levels.

Keywords: Heterogeneity; Latent classes; Latent variables; Mathematics achievement; Random effects

1. Introduction

Multilevel regression analysis has become a standard analysis tool for understanding unobserved heterogeneity in relationships between variables that are measured on individuals clustered within higher order units (Goldstein, 2003; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999). The heterogeneity is expressed in terms of random intercepts and slopes, i.e. continuous latent variables that vary between clusters. This paper argues for the need also to consider unobserved heterogeneity that represents qualitatively different relationships. This can be captured by categorical latent variables, i.e. latent classes. Such modelling is the domain of finite mixture analysis (McLachlan and Peel, 2000). For a review of finite mixtures applied to single-level regression analysis, see Wedel and DeSarbo (1993). In recent years several developments of multilevel finite mixture modelling have been made (Vermunt, 2003; Asparouhov and Muthén, 2008; Muthén and Asparouhov, 2008), but the topic still needs much further exploration.

This paper aims to shed light on the applicability of two-level regression mixture analysis through analyses of simulated and real data. The real data analyses investigate gender differences in mathematics achievement from the US National Education Longitudinal Survey (NELS) (US Department of Education, 1988). The analyses consider gender differences in selection mechanisms for mathematics training. A general two-level mixture model is used that encompasses

Address for correspondence: Bengt Muthén, Graduate School of Education, University of California, Los Angeles, 2023 Moore Hall, Mailbox 951521, Los Angeles, CA 90095-1521, USA.
E-mail: bmuthen@ucla.edu

both level 1 and level 2 mixtures. The 1988 NELS analyses show that the conventional two-level regression model is outperformed by the mixture analysis model. It is shown that level 1 heterogeneity in the form of latent classes can be mistaken for level 2 heterogeneity in the form of the random effects that are used in conventional two-level regression analysis. All analyses are carried out by using the Mplus program (Muthén and Muthén, 2008) and scripts of Mplus input code are available from the first author.

The outline of this paper is as follows. Section 2 presents prototypical examples that use artificial data to illustrate key differences between single-class and mixture models. Section 3 describes analyses using an educational data set. Section 4 concludes.

2. Prototypical examples

This section discusses examples of two prototypical mixture models. First, some basic ideas are illustrated by using a single-level regression mixture model. Second, a two-level structure is added.

2.1. Single-level regression mixture modelling

Consider a regression mixture model where the intercept and slope of a linear regression of a continuous variable y on a covariate x for individual i vary across the latent classes of a categorical latent variable C with K categories or latent classes labelled $c = 1, 2, \dots, K$,

$$y_{i|C_i=c} = \beta_{0c} + \beta_{1c} x_i + r_i, \quad (1)$$

where the residual $r_i \sim N(0, \theta_c)$. For parsimony, the residual variance θ_c is often held class invariant.

The probability of latent class membership varies as a multinomial logistic regression function of a covariate z ,

$$P(C_i = c | z_i) = \frac{\exp(a_c + b_c z_i)}{\sum_{s=1}^K \exp(a_s + b_s z_i)}, \quad (2)$$

where for the last class K the standardization $a_K = 0$ and $b_K = 0$ is used to designate this as a reference class. It should be noted that equations (1) and (2) generally may have different covariates x and z . In the examples that are considered below and in many applications, however, they are the same and z in equation (2) could be replaced with x . The mixture of normals provides a flexible representation of the distribution of y conditional on x and z .

As a simple example, assume that there are two latent classes and that y is the score on a mathematics test and x a binary variable representing gender. Consider data that are generated as follows. Let $x = 0$ represent males and $x = 1$ represent females. The β_0 -values for the two classes are chosen as 3 and 0, whereas the β_1 -values are -1.1 and 0.5 . The a -value is 0 and the b -value is 1. The covariate x has mean 0 and variance 1. The residual variance of y is 1. This gives a y -variance that is close to 2 in the high class and 1.25 in the low class. The class means of y for males are 3 and 0 respectively and for females 1.9 and 0.5. The class separation with respect to the y -mean is 3 standard deviations for males and 1.4 standard deviations for females. This scenario has features that are similar to what will be seen in the real data analyses to follow, although they are more accentuated here.

Figs 1(a)–1(d) show histograms for a sample of size 10000 generated from this model where in the latent class with a low score in mathematics (Figs 1(a) and 1(c)) females (Fig. 1(a)) do

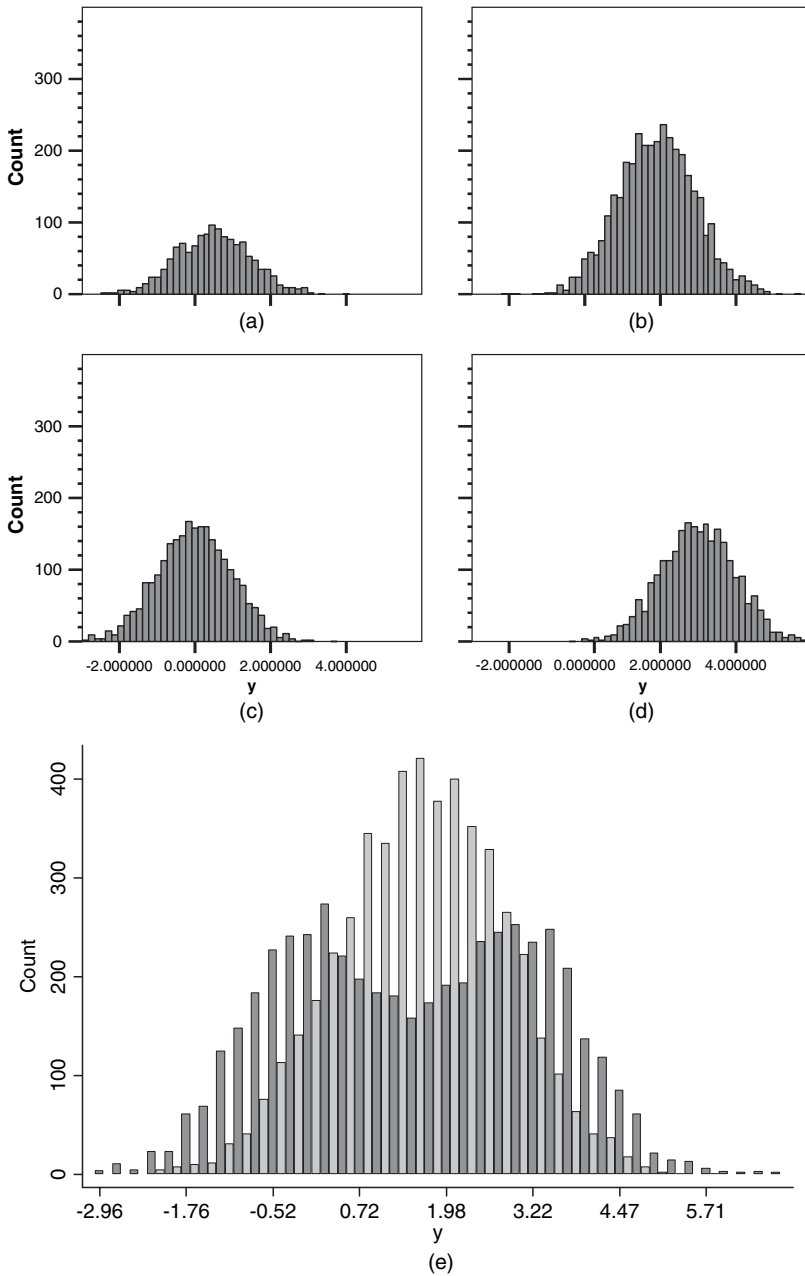


Fig. 1. Simulated two-class data for males and females: (a) low class, females; (b) high class, females; (c) low class, males; (d) high class, males; (e) histograms for the observed distributions (■, males; □, females)

better than males, whereas in the latent class with a high mathematics score males do better than females. The logistic regression model part (2) specifies that females are more likely to be in the high scoring latent class.

Fig. 1(e) shows the histograms for the observed distributions, i.e. ignoring latent class membership. It is seen that the mean is about the same for males and females, thereby hiding the

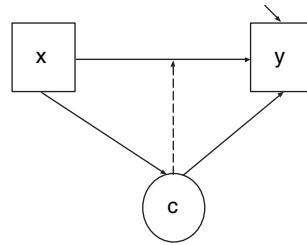


Fig. 2. Model diagram for single-level regression mixture analysis

gender level difference that is seen in each latent class. The histogram for males shows a bimodal distribution indicative of a two-class mixture, whereas for females the two classes generate a slightly skewed distribution. In this case the mathematics score means for the two classes are 3 standard deviations apart for males but only about half that for females. If the mathematics score means were closer together, bimodality would disappear and the distributions would look close to normal, thereby not suggesting a mixture.

Fig. 2 shows a diagram of the regression mixture model in equations (1) and (2) with observed variables in squares and the latent class variable in a circle. Full arrows represent regression relations. The arrows from x and c to y represent a linear regression of y on these two covariates. The influence of c on y is that of a dummy variable affecting the intercept and therefore the mean of y . The arrow from x to c indicates a logistic regression. The broken arrows from c to the regression of y on x indicate that the slope varies across the latent classes. The full arrow pointing to y and not originating from another variable represents the residual r in model (1). Further background on how to interpret model diagrams of this kind is given in Muthén and Muthén (2008).

The model that was used for the above simulated data can be understood as follows for the case of a continuous covariate x . An increasing value of the covariate x increases the probability of being in the latent class with the highest y -mean (class 1). Within each class, x has a further influence on y so as x increases y decreases for the high class (class 1) and y increases for the low class (class 2). In this sense, x has both an indirect influence on y via the latent class mediator variable c and a direct influence on y (for an overview of mediational modelling, see MacKinnon (2008)). It is interesting to note that the direct x influence on the dependent variable y is different even in sign in the two latent classes. The latent classes are to a large extent distinguished by the level of the dependent variable, but also the sign of the relationship which varies with the level. In this way, the regression mixture model shows a modelling flexibility that is not achieved in conventional regression. The mixture may be simply viewed as a tool for achieving this flexibility. In some situations, the latent classes of the mixture may also correspond to substantively meaningful clusters that represent distinct subpopulations of individuals, but this is not essential (for a discussion of the related concepts of direct and indirect uses of mixture modelling, see Titterington *et al.* (1985)). Using more than two classes provides further flexibility. In summary, the regression mixture model provides a modelling tool that is very flexible in describing situations that are more complex than those which are captured by conventional regression. For further details on single-level regression mixture modelling, see for example DeSarbo and Cron (1988).

2.2. Two-level regression mixture modelling

A conventional two-level regression model allows for variation in the dependent variable due to both level 1 and level 2 sources. Conditional on covariates, the level 1 source of variation

is typically confined to a residual variance that is common to all subjects, whereas the level 2 source of variation is expressed in terms of level 2 variation in random intercepts and random slopes. The level 1 variation is typically much larger than the level 2 variation, indicating greater unobserved heterogeneity on the individual level. Variation across subjects in the level 1 residual variance is sometimes found (see, for example Goldstein (2003)). It is argued here, however, that a potentially large source of unobserved heterogeneity resides in variation of the regression coefficients between groups of individuals sharing similar but unobserved background characteristics. It seems possible that this is quite common owing to heterogeneous subpopulations in general population surveys. Such heterogeneity is captured by level 1 latent classes.

These ideas can be formalized as follows. Consider a two-level regression mixture model where the random intercept and slope of a linear regression of a continuous variable y on a covariate x for individual i in cluster j vary across the latent classes of an individual level latent class variable C with K categories labelled $c = 1, 2, \dots, K$,

$$y_{ij|C_{ij}=c} = \beta_{0c} + \beta_{1c}x_{ij} + r_{ij}, \tag{3}$$

where the residual $r_{ij} \sim N(0, \theta_c)$. Here, a single covariate is used for simplicity of illustration, but further covariates can clearly be added. The probability of latent class membership varies as a two-level multinomial logistic regression function of a covariate z ,

$$P(C_{ij} = c | z_{ij}) = \frac{\exp(a_c + b_c z_{ij})}{\sum_{s=1}^K \exp(a_s + b_s z_{ij})}. \tag{4}$$

The corresponding level 2 equations are

$$\beta_{0c} = \gamma_{00c} + \gamma_{01c}w_{0j} + u_{0j}, \tag{5}$$

$$\beta_{1c} = \gamma_{10c} + \gamma_{11c}w_{1j} + u_{1j}, \tag{6}$$

$$a_c = \gamma_{20c} + \gamma_{21c}w_{2j} + u_{2c}. \tag{7}$$

With K categories for the latent class variable there are $K - 1$ equations (7). Here, w_{0j} , w_{1j} and w_{2j} are level 2 covariates and the residuals u_{0j} , u_{1j} and u_{2c} are $(2 + K - 1)$ -variate normally distributed with means 0 and covariance matrix Θ_2 and are independent of r_{ij} . As before, in many cases $z = x$ in equation (4). Also, the level 2 covariates in equations (5)–(7) may be the same as is the case in the examples considered below, where there is a common $w_j = w_{0j} = w_{1j} = w_{2j}$. For some or all of these relations there may be no covariates at all.

Fig. 3 shows a diagram of a special case of a two-level regression mixture model. To make it more concrete, the earlier two-class example of gender influence on performance in mathematics may be expanded to consider students observed within schools. The filled circles at level 1 indicate that the y - and c -regressions have school varying random intercepts, β_{0c} and a_c respectively, as in equations (3) and (4). Ignoring for the moment the latent class aspect, the meaning of a random intercept for y is well known. The random intercept for c , however, is for a multinomial logistic regression and therefore captures schools' influences on the students' probabilities to be in a certain latent class. At level 2 these random intercepts are continuous latent variables which are influenced by the level 2 covariate w as in equations (5)–(7), but now with the common covariate w . Let w be average socio-economic status (SES) of students in the school and have a positive influence γ_{21c} on a_c for the latent class with high mathematics score, $c = 1$ (a_c is standardized at 0 for $c = 2$ in line with conventional multinomial logistic regression).

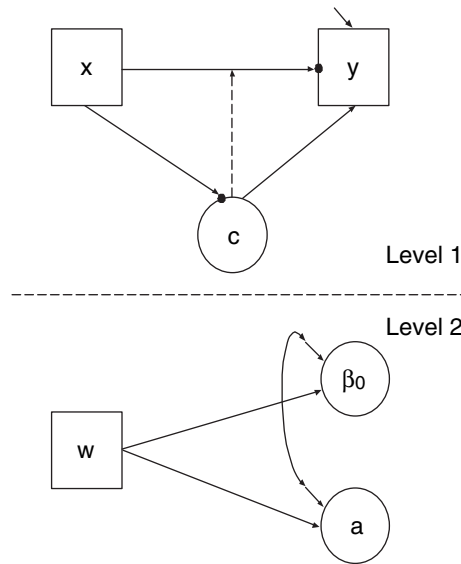


Fig. 3. Model diagram for two-level regression mixture analysis

In this way, increasing school SES increases a_{1j} and therefore the probability of being in the latent class that is characterized by high mathematics score and a female disadvantage. Within each class, let γ_{01c} in equation (5) be positive so that increasing school SES increases β_{0cj} and therefore the student mathematics score.

Using the above example as background and drawing on the notation of equations (3)–(7), it is instructive to consider the different components of the specification of the intercept and slope for y that are given in equations (5) and (6). The intercepts and slopes for y in the Fig. 3 model are expressed as

$$\beta_{0cj} = \gamma_{00c} + \gamma_{01}w_j + u_{0j}, \tag{8}$$

$$\beta_{1cj} = \gamma_{10c}. \tag{9}$$

For the random intercept β_{0cj} , equation (8) shows that the intercept γ_{00c} varies across the latent classes of the level 1 variable c . This may account for the majority of the variation in the random intercept with some further variation across level 2 units captured by the terms $\gamma_{01}w_j + u_{0j}$. Here, the slope γ_{01} is not specified as varying across the latent classes of the level 1 variable c , which may be realistic given that the influence of w on the random intercept is a level 2 relationship. In contrast with the random intercept, equation (9) shows that the slope β_{1cj} is specified as not varying between level 2 units but only across the latent classes of the level 1 variable c . In other words, for the slope it is assumed that it is more important to account for level 1 heterogeneity than level 2 heterogeneity. This seems often to be so in real data.

It is instructive to generate data from the model of Fig. 3 and to analyse them by using the conventional (single-class) two-level regression model with random intercepts and slopes,

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + r_{ij}, \tag{10}$$

where the residuals r_{ij} are assumed independently and normally distributed with zero mean and variance θ , and

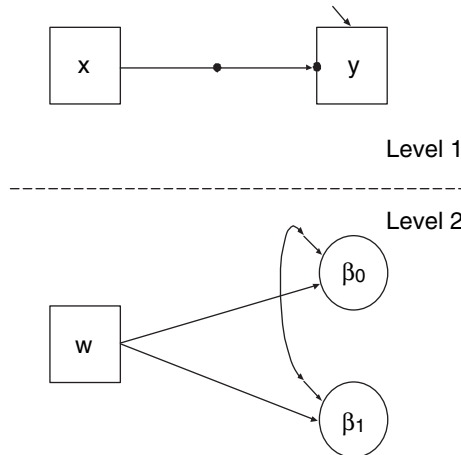


Fig. 4. Model diagram for conventional two-level regression analysis

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}, \tag{11}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}, \tag{12}$$

where the u_{0j} - and u_{1j} -residuals are assumed to be normally distributed with zero mean and covariance matrix Θ_2 . This conventional two-level model is shown in Fig. 4.

Data were generated for the two-level regression mixture model of Fig. 3 with 200 clusters each of size 15 for a total sample of $n = 3000$. The model of Fig. 3 has 11 parameters. The level 1 parameters were the same as those used above for the single-level regression mixture model. The level 2 parameters result in an intraclass correlation of 0.25 for the high class and 0.33 for the low class (the mean and variance of w were set at 0 and 1, $\gamma_{011} = 0.5$, $\gamma_{211} = 1$, the variance for the residual $u_{0j} = 0.5$ and the variance for the residual $u_{2j} = 0.5$). Using 100 replications, all parameter values as well as their standard errors were well recovered with good coverage.

The analysis of Fig. 3 data by using the conventional model of Fig. 4 and equations (10)–(12) points to interesting results that challenge the interpretations that are obtained by the conventional model. The conventional model has eight parameters. For each replication the log-likelihood value and the Bayesian information criterion BIC (Schwarz, 1978) were considerably worse for the conventional model than when using the correct Fig. 3 mixture model. This shows that the Fig. 3 model is clearly recognizable. The key result is that γ_{11} , the slope in the regression of the random slope on the level 2 covariate in equation (12), comes out negative and significant (not including the w -covariate shows a significant random-slope variance, so such a regression is natural to explore). In other words, a researcher would conclude that increasing school SES causes a reduction in the slope in the mathematics score regression on the variable female so that females are more disadvantaged relative to males in high SES schools. The incorrect attribution of the student level heterogeneity across the two latent classes to heterogeneity across schools is understandable and is probably common for the following reason.

The Fig. 3 model says that school SES has a positive influence on the random intercept for the logistic regression, which in turn increases the probability of being in the latent class that is characterized by a high score in mathematics and a female disadvantage (referred to as the high class) as opposed to the latent class with a low score in mathematics and a female advantage. This means that there is variation in the slope for females as school SES increases and the change is in the negative direction. But the conventional model makes the researcher believe that school

SES is the only cause of this gender phenomenon, whereas in reality it is caused by student level latent class membership which only partly has school level causes. In the simulated data, the parameter values are chosen so that the influence on the probability of being in the high class is the same for the student level covariate female as for a change of 1 standard deviation in the school level SES covariate.

Given that the probability of membership in the high latent class is higher for females than males, one possible scenario for the latent class phenomenon is that females choose more rigorous mathematics studies than males, but once they have done so they have a disadvantage relative to males. Other student level covariates can be sought that influence and explain the latent class membership, such as prior performance in mathematics. Concluding on this chain of events is quite different from the conclusion that is drawn from the conventional model. It is true that increasing school SES has an influence by increasing the probability of membership in the high class, but this influence is only half of the story. From an intervention point of view, the conventional model and the mixture model may lead to different decisions on what to manipulate. The conventional model ignores that membership in the high class is a student level decision. A school level intervention might be less effective than a student level intervention.

To conclude this section, it is instructive to turn the situation around and to use the mixture model to analyse data that are generated according to a conventional random-effect regression model. The numbers of students and schools are the same as in the previous data generation. To accentuate the school level variation in the random effects, a model was chosen with random-intercept variance that is three-quarters of the within-level residual variance and a random-slope variance that is one and a half times the within-level residual variance. It should be noted that the strong random-slope variation generates a non-normal dependent variable y which might lead a researcher to use a mixture of normal distributions. The level 1 residual variance of y is 1, the mean and variance of the covariate w are 0 and 1, $\gamma_{00} = 0$, $\gamma_{01} = 0.5$, the variance of u_{0j} is 0.5, $\gamma_{10} = 1$, $\gamma_{11} = 1$, the variance of u_{1j} is 0.5 and the u_0, u_1 covariance is 0. The conventional random-effect model has eight parameters whereas a two-class mixture model which is specified as in Fig. 3 has 11 parameters. For each replication the log-likelihood and BIC-values of the two-class mixture model were worse than those of the conventional (single-class) model, clearly rejecting the need for more than one class. This suggests that the mixture model is not likely to be found if the correct model does not have a mixture. Despite strong random-slope variation of the conventional model, generating a non-normal dependent variable y , the log-likelihood and BIC do not indicate a need for a mixture.

3. Real data example

In this section the conventional two-level regression model as well as two-level regression mixture models are applied to data from the 1988 NELS (US Department of Education, 1988). Mathematics achievement data from the 1988 grade 8 data are considered for a sample of 14217 students in 913 US schools. In the USA grade 8 is the ninth year of formal schooling and is typically referred to as middle school with students of age 13 years. The average number of students per school is 15.6. The intraclass correlation for the mathematics score is 0.13. Of the schools in the sample, 85% are public, 8% are private and 7% are Catholic. The analyses are merely illustrative of the methodology and use a limited set of variables similar to the well-known 'High school and beyond' example that was used as a key illustration in Raudenbush and Bryk (2002). The level 1 covariates are gender (*female*) and student SES (*stud_ses*). The level 2 covariates include the percentage of teachers with at least a Masters degree (*per_adva*), dummy variables

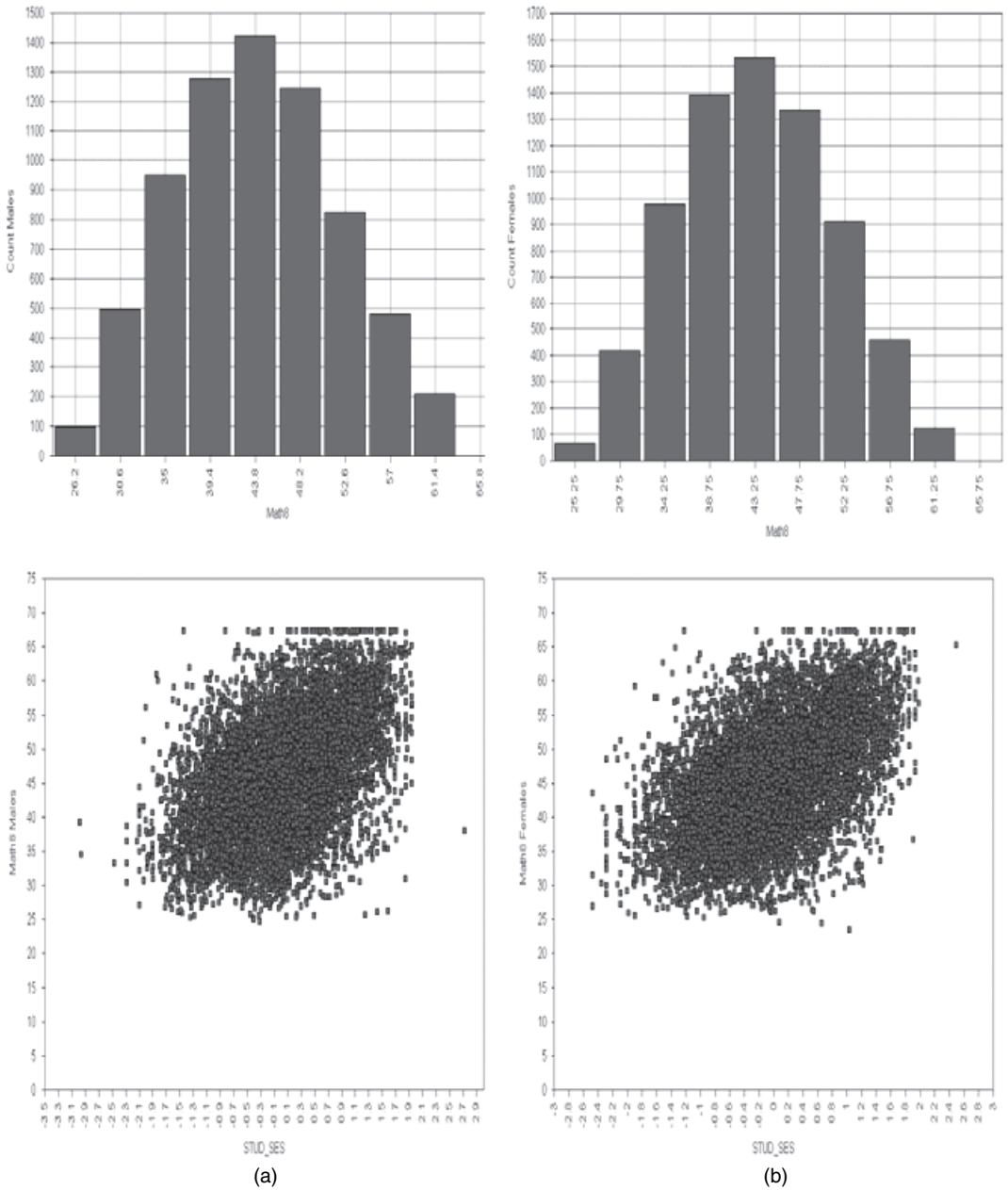


Fig. 5. Descriptive statistics for the NELS mathematics achievement data related to gender and student SES: (a) males; (b) females

for private and Catholic schools compared with public schools (*private* and *Catholic*), and the average SES in the school (*mean_ses*). The continuous covariates are grand mean centred. The histograms for the mathematics scores of males and females are shown in Fig. 5 and seem to indicate distributions that do not deviate far from normal. The non-normality that is seen in Fig. 1 is not apparent here. Also, no clear outlying groups of students are detected in the scatter plots of math on stud_ses.

All subsequent analyses were carried out by using maximum likelihood estimation in the Mplus program (Muthén and Muthén, 2008). The mixture modelling draws on the techniques that are described in Asparouhov and Muthén (2008) and Muthén and Asparouhov (2008).

3.1. Conventional analysis

The conventional two-level regression model applied to NELS mathematics achievement is

$$\text{math}_{ij} = \beta_{0j} + \beta_{1j} \text{female}_{ij} + \beta_{2j} \text{stud_ses}_{ij} + r_{ij}, \tag{13}$$

where the residuals r_{ij} are assumed independently normally distributed with zero mean and variance θ , and

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \text{per_adva} + \gamma_{02} \text{private} + \gamma_{03} \text{Catholic} + \gamma_{04} \text{mean_ses} + u_{0j}, \tag{14}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \text{per_adva} + \gamma_{12} \text{private} + \gamma_{13} \text{Catholic} + \gamma_{14} \text{mean_ses} + u_{1j}, \tag{15}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} \text{per_adva} + \gamma_{22} \text{private} + \gamma_{23} \text{Catholic} + \gamma_{24} \text{mean_ses} + u_{2j}, \tag{16}$$

where the u_{0j} -, u_{1j} - and u_{2j} -residuals are assumed to be normally distributed with zero means and covariance matrix Θ_2 .

Table 1 gives the estimates and standard errors for the conventional model with a random intercept and two random slopes. The grand mean centring implies that β_{0j} refers to males

Table 1. NELS mathematics analysis: estimates from conventional (single-class) two-level regression analysis

<i>Level</i>	<i>Parameter</i>	<i>Estimate</i>	<i>Standard error</i>	<i>Z-ratio</i>	<i>P-value</i>
1	Residual variance	50.183	0.757	66.291	0.000
2	Regression of mathematics intercept β_{0j}				
	on per_adva, γ_{01}	0.274	0.570	0.481	0.631
	on private, γ_{02}	1.867	0.846	2.207	0.027
	on Catholic, γ_{03}	-0.591	0.503	-1.174	0.240
	on mean_ses, γ_{04}	3.488	0.313	11.135	0.000
2	Regression of female slope β_{1j}				
	on per_adva, γ_{11}	0.139	0.643	0.217	0.828
	on private, γ_{12}	-0.510	0.683	-0.746	0.456
	on Catholic, γ_{13}	0.028	0.583	0.048	0.962
	on mean_ses, γ_{14}	0.014	0.318	0.044	0.965
2	Regression of ses slope β_{2j}				
	on per_adva, γ_{21}	0.906	0.394	2.298	0.022
	on private, γ_{22}	-1.927	0.509	-3.786	0.000
	on Catholic, γ_{23}	-1.548	0.403	-3.841	0.000
	on mean_ses, γ_{24}	0.829	0.211	3.936	0.000
2	Intercept for mathematics intercept, γ_{00}	45.680	0.145	316.013	0.000
	for female slope, γ_{10}	-0.225	0.155	-1.447	0.148
	for ses slope, γ_{20}	3.737	0.108	34.764	0.000
2	Residual variance for mathematics intercept	6.687	0.667	10.024	0.000
	for female slope	3.242	0.776	4.177	0.000
	for ses slope	0.014	0.316	0.044	0.965
2	Residual covariance, mathematics intercept, female slope	-2.430	0.603	-4.028	0.000
	mathematics intercept, ses slope	0.195	0.274	0.712	0.477

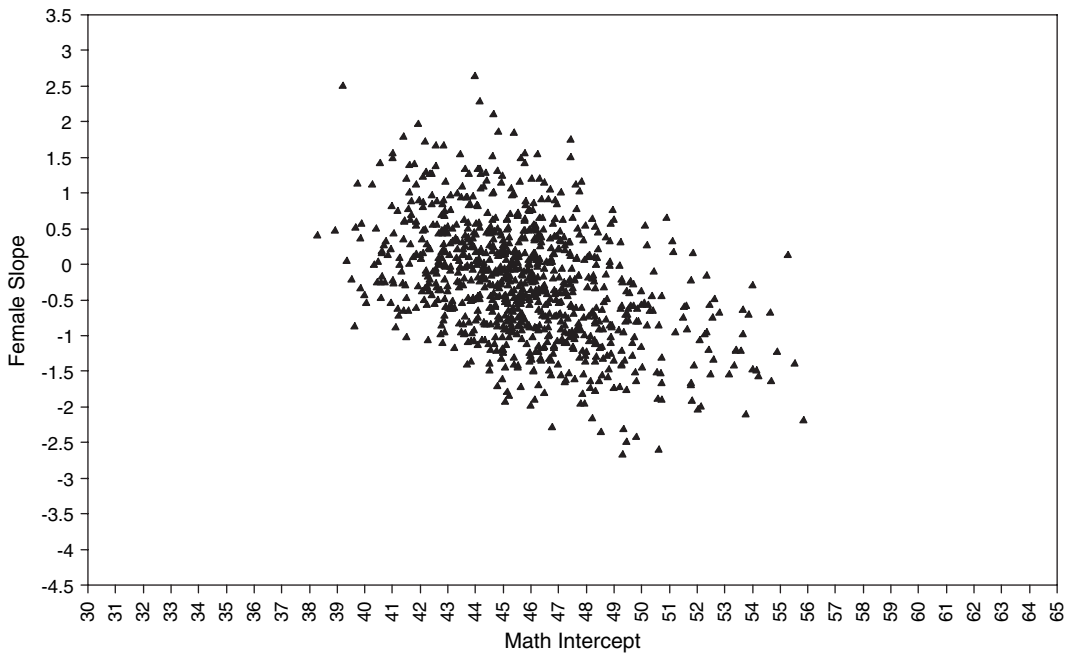


Fig. 6. Plot of school-estimated slopes for females related to estimated mathematics intercepts for the NELS by using conventional two-level regression analysis

at an average *stud_ses*-value and that γ_{00} refers to public schools at school average values of *per_adv* and *mean_ses*. It is seen that the random intercept β_{0j} is significantly influenced by private (positive) and *mean_ses* (positive). The random slope for female has no significant school level predictor. The average of the slope for females for public schools is -0.225 , which is not a significant gender effect. The random slope for student SES is significantly influenced by the school level covariates *per_adv* (positive), private (negative), Catholic (negative) and *mean_ses* (positive). The average slope for *stud_ses* is estimated as 3.737. Taken together, this implies that private and Catholic schools reduce the positive effect of student SES on mathematics and are in this sense more ‘egalitarian’.

Fig. 6 shows the estimated random slopes for the covariate female plotted against the estimated random intercepts for math for each of the 913 schools in the sample (because the level 1 gender covariate is scored as 0 or 1 for male or female the random intercept represents the score for males at public schools with average school values on *per_adv* and *mean_ses*). These are expected *a posteriori* estimates drawing on normality assumed for the level 2 residuals in the model. Fig. 6 illustrates the negative relationship between these random effects. The slopes for females include both negative and positive values representing mathematics disadvantage *versus* advantage relative to males. In high performing schools females have a disadvantage relative to males in terms of mathematics score whereas in low performing schools females have an advantage. Females have an advantage in 38% of the schools.

The aim of the alternative mixture modelling to be presented in the next sections is to investigate the need to account for unobserved level 1 heterogeneity. Because of this, it is of interest to extend the conventional model to investigate level 1 heterogeneity more fully in the form of a level 1 residual variance varying across level 1 covariate values (see, for example, Goldstein (2003)) and across the three school types, public, Catholic and private (these exten-

sions were suggested by a reviewer). The level 1 residual variance relationship to level 1 covariates was investigated by adding the following heterogeneous residual variance model to equations (13)–(16):

$$r_{ij} = N(0, \kappa_0 + \kappa_1 \text{female}_{ij} + \kappa_2 \text{stud_ses}_{ij}), \tag{17}$$

where κ_0 , κ_1 and κ_2 are parameters to be estimated. The κ_1 - and κ_2 -estimates were found to be significant, indicating a somewhat larger level 1 residual variance for males than females and a small increase with increasing *stud_ses*. Allowing the level 1 residual variance to vary as a function of school type resulted in significantly different residual variances with the smallest residual variance for private schools, followed by Catholic schools, and with public schools obtaining the largest value. The log-likelihood values for these two models are reported later as model 1b and model 1c respectively, in the Table 2 summary for all models.

3.2. Mixture analysis

In this section, two-level regression mixture modelling is applied to the NELS mathematics achievement data. Three mixture models are explored: model A, model B and model C. The key model B will be presented not only in terms of formulae but also in a model figure for readers who are not accustomed to complex statistical equations.

3.2.1. Mixture modelling with level 1 latent class variable: models A and B

Fig. 7 shows the two-level regression mixture model diagram (referred to as model B below). The diagram uses the conventions that were described for Figs 2–4. It shows that unobserved heterogeneity in the mathematics intercept and slopes is allowed for not only among schools, as in the conventional model, but also among students. For students, the heterogeneity is expressed as latent classes and for schools the heterogeneity is expressed as random coefficients varying between schools. In this way, the model captures heterogeneity by using both categorical and continuous latent variables.

Fig. 7 assumes three latent classes ($K = 3$). At level 1 the arrows pointing to c from the covariates *female* and *stud_ses* represent the multinomial logistic regression of student latent class membership. The two filled circles for c represent the two random intercepts for $c = 1$ and $c = 2$ (the third class is the reference class). The three filled circles in the linear regression of *math* on *female* and *stud_ses* represent the random intercept and the two random slopes. At level 2 the random effects are related to the covariates *per_adva*, *private*, *Catholic* and *mean_ses* by using linear regressions. The absence of a residual arrow for *s_ses* indicates that the residual variance for *s_ses* is held at 0 because it was found insignificant. The other residuals are correlated.

Formally, the Fig. 7 two-level regression mixture model is expressed as follows. Consider the random intercept and slopes of the linear regression of *math* regressed on the covariates *female* and *stud_ses* for individual i in school j varying across the latent classes of an individual level categorical latent variable C with K categories or latent classes,

$$\text{math}_{ij|C_{ij}=c} = \beta_{0cj} + \beta_{1cj} \text{female}_{ij} + \beta_{2cj} \text{stud_ses}_{ij} + r_{ij}, \tag{18}$$

where the residual $r_{ij} \sim N(0, \theta)$ and independently distributed. The probability of latent class membership varies as a two-level multinomial logistic regression function,

$$P(C_{ij} = c | \text{female}_{ij}, \text{stud_ses}_{ij}) = \frac{\exp(a_{cj} + b_{1c} \text{female}_{ij} + b_{2c} \text{stud_ses}_{ij})}{\sum_{s=1}^K \exp(a_{sj} + b_{1s} \text{female}_{ij} + b_{2s} \text{stud_ses}_{ij})}. \tag{19}$$

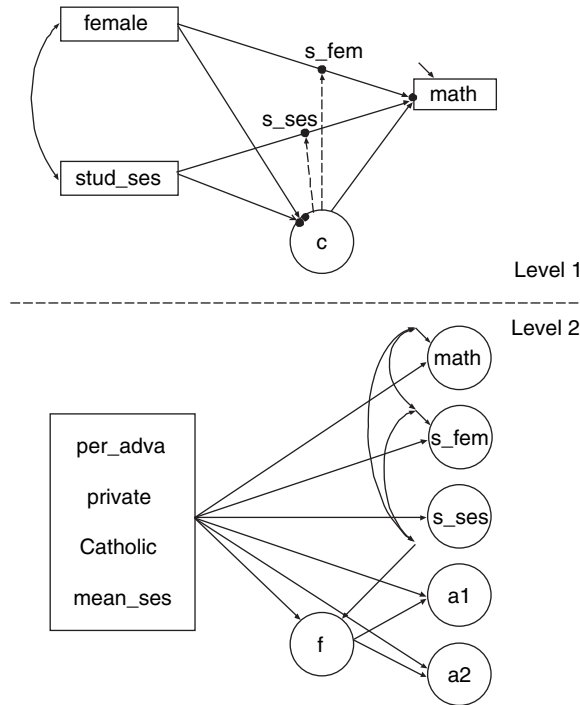


Fig. 7. Model diagram for two-level regression mixture analysis of NELS mathematics achievement (model B)

The corresponding level 2 equations are written as functions of the level 2 covariates *per_adva*, *private*, *Catholic* and *mean_ses*:

$$\beta_{0cj|C_{ij}=c} = \gamma_{00c} + \gamma_{01c} \text{ per_adva}_j + \gamma_{02c} \text{ private}_j + \gamma_{03c} \text{ Catholic}_j + \gamma_{04c} \text{ mean_ses}_j + u_{0j}, \quad (20)$$

$$\beta_{1cj|C_{ij}=c} = \gamma_{10c} + \gamma_{11c} \text{ per_adva}_j + \gamma_{12c} \text{ private}_j + \gamma_{13c} \text{ Catholic}_j + \gamma_{14c} \text{ mean_ses}_j + u_{1j}, \quad (21)$$

$$\beta_{2cj|C_{ij}=c} = \gamma_{20c} + \gamma_{21c} \text{ per_adva}_j + \gamma_{22c} \text{ private}_j + \gamma_{23c} \text{ Catholic}_j + \gamma_{24c} \text{ mean_ses}_j + u_{2j}, \quad (22)$$

$$a_{cj} = \gamma_{30c} + \gamma_{31c} \text{ per_adva}_j + \gamma_{32c} \text{ private}_j + \gamma_{33c} \text{ Catholic}_j + \gamma_{34c} \text{ mean_ses}_j + u_{3j}, \quad (23)$$

where the last equation represents $K - 1$ random-effects regressions and the residuals u_{0j} , u_{1j} , u_{2j} and u_{3j} are $(3 + K - 1)$ -variate normally distributed with means 0 and covariance matrix Θ_2 and are independent of r_{ij} .

With K classes, there are $K - 1$ random effects u_{3j} in equation (23). In line with Vermunt (2003) the dimensionality of the multinomial random intercepts can be reduced by using a continuous factor f to represent their variation and covariation. The dimension reduction is important in the maximum likelihood estimation because each random effect corresponds to one dimension of numerical integration, where an increasing number of integration dimensions leads to heavy and ultimately intractable computations. With three latent classes ($K = 3$), the two multinomial random intercepts a_{1j} and a_{2j} can be expressed in terms of f_j as

$$a_{1j} = f_j, \quad (24)$$

$$a_{2j} = \lambda f_j. \quad (25)$$

The variances and covariance of the two intercepts given the level 2 covariates are thereby parameterized in terms of the parameter λ and the residual variance of the factor f_j . In the NELS application below, the three classes can be seen as ordered on a dimension from high to low performance in mathematics in which case the single-factor simplification seems reasonable in that the two random effects are likely to be highly correlated but have different variances. The factor f_j is in turn linearly regressed on the level 2 covariates. To allow the two multinomial random intercepts to have different relations to the level 2 covariates, the covariates are allowed to have a direct influence on one of the intercepts (here the second). In equations (24) and (25) there are no residuals for the two multinomial random intercepts in their linear regression on f , reducing two dimensions to one. This mixture model will be referred to as model B.

A simplified version of this model, model A, expresses the heterogeneity in the slopes for the covariates female and stud_ses by using only the student level latent classes, not allowing further random variation in these slopes across schools, i.e. $u_{1j} = 0$ and $u_{2j} = 0$. This leads to two-dimensional integration corresponding to the mathematics intercept and the factor f . This model is interesting in that it completely replaces the between-school variation in the random slopes with between-student variation. The random intercept, however, retains the variation between both students and schools given that this variation is typically much more pronounced.

3.2.2. Mixture modelling with level 1 and level 2 latent class variable: model C

Consider the two-level regression mixture model where the random intercept and slope of the linear regression of the mathematics achievement score math regressed on the covariates female and stud_ses for individual i in school j vary across the categories of an individual level latent class variable C with K categories as well as across a school level latent class variable D with L categories,

$$\text{math}_{ij|C_{ij}=c, D_j=d} = \beta_{0cdj} + \beta_{1cdj} \text{female}_{ij} + \beta_{2cdj} \text{stud_ses}_{ij} + r_{ij}, \tag{26}$$

where the residuals $r_{ij} \sim N(0, \theta)$ and independently distributed. The probability of latent class membership for the individual level latent class variable varies as a two-level multinomial logistic regression function:

$$P(C_{ij} = c | D_j = d, \text{female}_{ij}, \text{stud_ses}_{ij}) = L_{ij},$$

$$L_{ij} = \frac{\exp(a_{cj} + b_{1c} \text{female}_{ij} + b_{2c} \text{stud_ses}_{ij})}{\sum_{s=1}^K \exp(a_{sj} + b_{1s} \text{female}_{ij} + b_{2s} \text{stud_ses}_{ij})}. \tag{27}$$

The corresponding level 2 equations are written as functions of the level 2 covariates per_adva, private, Catholic and mean_ses, with variation in coefficients across the latent classes of both C and D ,

$$\beta_{0c|D_j=d} = \gamma_{00cd} + \gamma_{01cd} \text{per_adva}_j + \gamma_{02cd} \text{private}_j + \gamma_{03cd} \text{Catholic}_j + \gamma_{04cd} \text{mean_ses}_j + u_{0j}, \tag{28}$$

$$\beta_{1c|D_j=d} = \gamma_{10cd} + \gamma_{11cd} \text{per_adva}_j + \gamma_{12cd} \text{private}_j + \gamma_{13cd} \text{Catholic}_j + \gamma_{14cd} \text{mean_ses}_j + u_{1j}, \tag{29}$$

$$\beta_{2c|D_j=d} = \gamma_{20cd} + \gamma_{21cd} \text{per_adva}_j + \gamma_{22cd} \text{private}_j + \gamma_{23cd} \text{Catholic}_j + \gamma_{24cd} \text{mean_ses}_j + u_{2j}, \tag{30}$$

$$a_{cj|D_j=d} = \gamma_{30c} + \gamma_{31c} \text{per_adva}_j + \gamma_{32c} \text{private}_j + \gamma_{33c} \text{Catholic}_j + \gamma_{4c} \text{mean_ses}_j + u_{3j}, \quad (31)$$

where the last equation represents $K - 1$ random-effects regressions and the residuals u_{0j}, u_{1j}, u_{2j} and u_{3j} are $(3 + K - 1)$ -variate normally distributed with means 0 and covariance matrix Θ_2 and are independent of r_{ij} . The same dimension reduction as in equations (24) and (25) is carried out here. The model is completed by expressing the probability of latent class membership for the school level latent class variable D as the multinomial logistic regression,

$$P(D_j = d | \text{per_adva}_j, \text{private}_j, \text{Catholic}_j, \text{mean_ses}_j) = L_j, \\ L_j = \frac{\exp(a_d + b_{1d} \text{per_adva}_j + b_{2d} \text{private}_j + b_{3d} \text{Catholic}_j + b_{4d} \text{mean_ses}_j)}{\sum_{s=1}^L \exp(a_s + b_{1s} \text{per_adva}_j + b_{2s} \text{private}_j + b_{3s} \text{Catholic}_j + b_{4s} \text{mean_ses}_j)}. \quad (32)$$

3.3. National Education Longitudinal Survey model comparisons

For comparison of fit of models that have the same number of classes and are nested, the usual likelihood ratio and Wald χ^2 -tests will be used. Likelihood ratio comparisons of models with different numbers of classes violate the requirement of not having parameters on the border of the admissible parameter space in the more restricted model because of zero-probability parameters. Deciding on the number of classes is instead typically accomplished by the Bayesian information criterion BIC (Schwarz, 1978; Kass and Raftery, 1993),

$$\text{BIC} = -2 \log(L) + r \log(n), \quad (33)$$

where r is the number of free parameters in the model and n is the sample size. The lower the BIC-value, the better the model. The number of classes is increased until a BIC minimum is found.

Table 2 reports the log-likelihood, number of parameters and BIC for the models that were considered. The conventional model is referred to as model 1 in Table 2, whereas the extension to heterogeneous level 1 residual variance as a function of level 1 covariates is labelled model 1b and the extension to heterogeneous level 1 residual variance as a function of level 2 school type is labelled model 1c. Among the 1-class models, model 1c has the best (lowest) BIC-value. It is clear, however, that the BIC improvement for model 1c relative to model 1 is considerably

Table 2. NELS mathematics analysis: log-likelihood and BIC comparisons of models

Model	Number of classes	Log-likelihood	Number of parameters	BIC
1	1	-48519	21	97240
1b	1	-48507	23	97234
1c	1	-48486	23	97192
A2	2	-48358	20	96907
A3	3	-48288	28	96844
A4	4	-48255	35	96845
B2	2	-48337	32	96981
B3	3	-48258	43	96928
B4	4	-48208	54	96933
C23	2, 3	-48211	61	97004

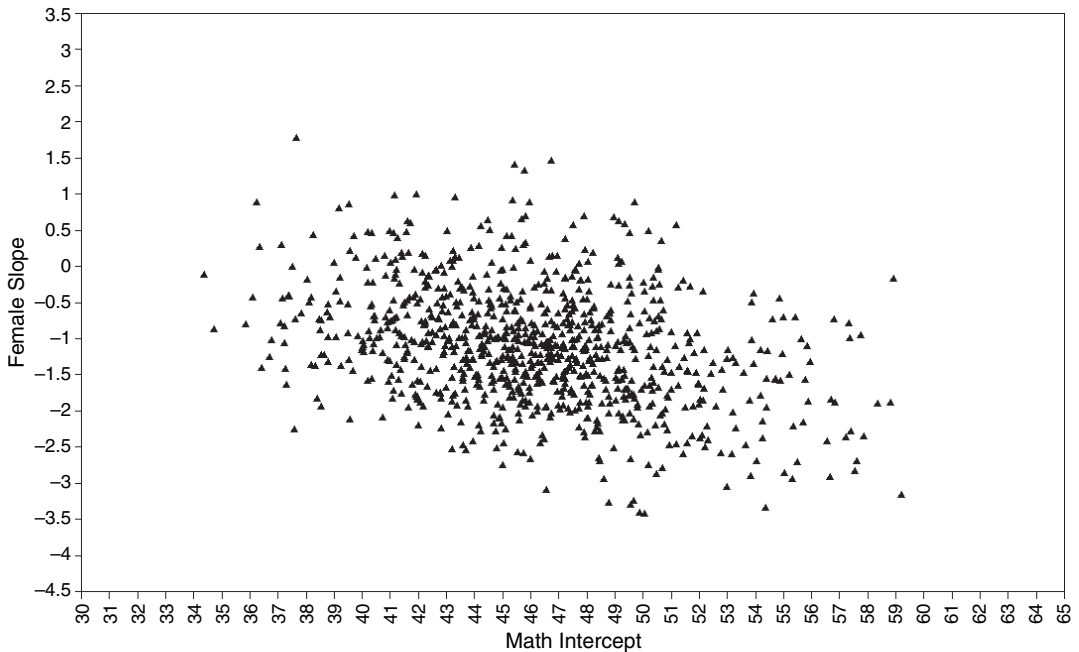


Fig. 8. School-estimated slopes for females related to estimated mathematics intercepts for the NELS by using two-level regression mixture analysis

smaller than the BIC improvement of the 2-class model A2 relative to model 1. Model A2 uses fewer parameters than model 1c but has a better log-likelihood and BIC, illustrating the importance of allowing for level 1 heterogeneity in terms of mixtures with coefficients varying across the latent classes. Model A3 with three latent classes has only five more parameters than model 1c but has a considerably lower BIC. Comparing the three versions of model A, model A3 is found preferable on the basis of BIC.

The results for the more general model of Fig. 7 are listed under model B in Table 2. Model B allows slope heterogeneity among both students and schools. For this type of model, BIC suggests that three latent classes are optimal. Although model B3 has a worse BIC-value than model A3, these two models have the same number of classes and can therefore be compared in a conventional way. The Wald test significance of the extra parameters of model B3 shows that model B3 is preferable to model A3. Model C23 adds two latent classes at level 2 to model B3, but BIC indicates that this school level mixture is not needed for these data.

The conclusions that are drawn from model B3 are quite different from those of the conventional model 1. Fig. 8 shows a scatter plot of the estimated slopes for females related to the mathematics intercepts for the 913 schools. Fig. 8 is based on the model 3 posterior estimates for the random intercept and slopes. Because these random effects have within-level variation over the latent classes the estimates for the random effects in the three possible level 1 classes are different and therefore the posterior estimates for students within the same school can differ. In addition each student can be partially classified in two or more classes with varying class membership probabilities, which leads to further variation in the random-effect estimates. Fig. 8 shows the scatter plot for the school average random intercept and the school average gender random slope. Fig. 8 may be compared with the corresponding Fig. 6 for model 1. Whereas for model 1 38% of the schools showed an advantage for females over males, in model B3 this is reduced to 9%.

The estimated model B3 may be discussed by using Fig. 7. At level 1 the influence of the three classes of c on mathematics and on the regression of mathematics on female and stud_ses is characterized as follows. The three latent classes are ordered from high to low mathematics achievement intercept (γ_{00}): 57.2 (class 1, 22%), 47.7 (class 2, 46%) and 37.4 (class 3, 32%). The mean differences correspond to a little more than 1 standard deviation of the mathematics score. The means of the slopes for females for these three classes evaluated for public schools (γ_{10}) all indicate a disadvantage for females: -2.0 , -1.3 and -0.2 . Here, the intercepts for class 1 and 2 are significantly different from 0 but the mean for class 3 is not. The significant gender differences are in contrast with the non-significance that is found in the conventional model where no latent class distinction was made. The level 1 multinomial logistic regression of latent class membership on the level 1 covariates shows that both female and stud_ses have significant positive influence on the probability of being in the high mathematics scoring class 1. In terms of gender effects, the results are similar to what was shown in Fig. 1.

At level 2 of model B3 the mathematics intercept is not significantly predicted by any of the level 2 covariates. For the conventional model the mathematics intercept had a positive significant influence from the level 2 covariates private and mean_ses (shown earlier in Table 1). For model B3, however, the mathematics intercept regression variation is expressed differently, being both between and within latent classes. The non-significance refers to the within-class variation. The level 2 covariates can also have an influence on performance in mathematics via the random intercepts for the multinomial logistic regression. Model B3 shows such an effect as a significant positive effect of mean_ses on the factor f . Unlike the conventional model, there is, however, not a significant effect of the sector covariate private. As in the conventional model, the model B3 level 2 regression of the slope for females on the level 2 covariates shows no significant effects. The residual variance for the slope for females is 3.1 with standard error 0.7 supporting the need for model B3 over the simpler model A3. For the random slope for stud_ses there is a significant negative influence from the sector covariate Catholic. Unlike the conventional model, however, there is no significant effect of private. In this way, the conclusions regarding egalitarian effects are different between the two models. Also, model B3 does not show a significant effect of mean_ses on the stud_ses slope. In conclusion, the conventional model and the mixture model provide different conclusions regarding key aspects of the analyses. The mixture model has better log-likelihood and BIC-values and can therefore be said to be closer to the data. This suggests that the mixture model results are more trustworthy.

4. Conclusions

The two-level regression mixture analyses have shown that unobserved heterogeneity should not be presupposed to exist at level 2 at the expense of level 1. Both the simulated and the real data analyses have shown that level 1 heterogeneity in the form of latent classes can be mistaken for level 2 heterogeneity in the form of the random effects that are used in conventional two-level regression analysis. This suggests that mixture models have an important role to play in multilevel regression analyses. Mixture models allow heterogeneity to be investigated more fully, more correctly attributing different portions of the heterogeneity to the different levels.

NELS analyses of mathematics performance in grade 12 gave similar results to the results discussed for grade 8. Although the analyses that are shown here are mostly illustrative of the methods and not based on substantive causal theories, they may serve to stimulate further

investigations of the determinants of student level latent class membership. Such investigations may focus on the mathematics achievement development over grades 8–12 and also using information on mathematics courses taken during these grades.

A drawback of the mixture modelling is that the maximum likelihood computations are significantly more time consuming. First, computations are slow because mixture modelling in general exhibits multiple maxima of the likelihood so analyses that are based on many random starting value sets must be used to avoid a local maximum. Second, maximum likelihood estimation of mixture models with random effects calls for numerical integration, where each random effect with a non-zero level 2 variance contributes one dimension of integration. Model B involves three dimensions of integration. Taken together with the large sample size of $n = 14217$ this makes the computations slow. Model B has only one random slope and, with more random slopes, the computations would be even slower. The simpler model A involves only two dimensions of integration irrespective of the number of slopes and is therefore a more practical model. For these data, model B fits better but model A results were close to those of model B and may serve as an approximation. Model C uses both level 1 and level 2 latent classes and leads to even more heavy computations. A model A version of this may give more manageable computations. On the whole, however, the rapidly improving speed of computers and the increased availability of multiple processors are likely ultimately to make these timing concerns less relevant. As an alternative, Markov chain Monte Carlo analysis may be performed as in Goldstein and Browne (2002, 2005). Markov chain Monte Carlo sampling, however, comes with its own challenges in the context of complex latent variable modelling, including choice of parameterizations to ensure good mixing, label switching for latent classes and difficulties in determining identification status and convergence.

Acknowledgements

The research of the first author was supported under grant K02 AA 00230-01 from the National Institute on Alcohol Abuse and Alcoholism, by grant 1 R21 AA10948-01A1 from the National Institute on Alcohol Abuse and Alcoholism, by the National Institute of Mental Health under grant MH40859 and by grant P30 MH066247 from the National Institute on Drug Abuse and the National Institute of Mental Health. The work has benefited from discussions in the Prevention Science and Methodology Group, directed by Hendricks Brown. We thank Shaunna Clark and Michelle Conn for helpful research assistance.

References

- Asparouhov, T. and Muthén, B. (2008) Multilevel mixture models. In *Advances in Latent Variable Mixture Models* (eds G. R. Hancock and K. M. Samuelson), pp. 27–51. Charlotte: Information Age Publishing.
- DeSarbo, W. S. and Cron, W. L. (1988) A maximum likelihood methodology for clusterwise linear regression. *J. Class.*, **5**, 248–282.
- Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Arnold.
- Goldstein, H. and Browne, W. (2002) Multilevel factor analysis modelling using Markov chain Monte Carlo estimation. In *Latent Variable and Latent Structure Models* (eds G. Marcoulides and I. Moustaki), ch. 11, pp. 225–244. Mahwah: Erlbaum.
- Goldstein, H. and Browne, W. (2005) Multilevel factor analysis models for continuous and discrete data. In *Contemporary Psychometrics* (eds A. Maydeu-Olivares and J. J. McArdle), ch. 14, pp. 453–475. Mahwah: Erlbaum.
- Kass, R. E. and Raftery, A. E. (1993) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.
- MacKinnon, D. (2008) *Introduction to Statistical Mediation Analysis*. New York: Erlbaum.
- McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.
- Muthén, B. and Asparouhov, T. (2008) Growth mixture modeling: analysis with non-Gaussian random effects. In *Longitudinal Data Analysis* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), pp. 143–165. Boca Raton: Chapman and Hall–CRC.

- Muthén, B. and Muthén, L. (2008) *Mplus User's Guide*. Los Angeles: Muthén and Muthén.
- Raudenbush, S. W. and Bryk, A. S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn. Newbury Park: Sage.
- Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
- Snijders, T. and Bosker, R. (1999) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oakes: Sage.
- Titterton, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.
- US Department of Education (1988) National Education Longitudinal Study of 1988 (NELS:88), base-year and first follow-up surveys. National Center for Education Statistics, US Department of Education, Washington DC.
- Vermunt, J. (2003) Multilevel latent class models. *Sociol. Methodol.*, **33**, 213–239.
- Wedel, M. and DeSarbo, W. S. (1993) A review of recent developments in latent structure regression models. In *Advanced Methods of Marketing Research* (ed. R. Bagozzi), pp. 352–388. London: Blackwell.